S L A C
Stanford Linear Accelerator Center

THINKING INSIDE THE BOX

CHUCK BOEHEIM

Assistant Director

Scientific Computing and Computing Services

SLAC

SLAC-PUB-12986

Revised 11/29/2007

Work supported by the U.S. Department of Energy under contract number DE-AC02-76SF00515.

Introduction

In early 2007, SLAC was faced with a shortage of both electrical power and cooling in the main computer building, at the same time that the BaBar collaboration needed a new cluster of 250 batch machines installed. A number of different options were explored for the expansion. Provision of additional electrical power to the building was estimated to take one to two years, and cost several million dollars; additional cooling was even worse. Space in a Silicon Valley co-location facilities was reasonable on a one-year timescale, but broke even in costs by the end of three years, and were more expensive after that. There were also unresolved questions about the affects of additional latency from an offsite compute cluster to the onsite disk servers. The option of converting existing experimental hall space into computer space was estimated at one year, with uncertain availability. An option to aggressively replace several existing clusters with more power-efficient equipment was studied closely, but was disruptive to continued operations, expensive, and didn't provide any additional headroom. Finally, the installation of a Sun Project Blackbox (PBB) unit was selected as providing the capacity on a timescale of six months for a reasonable cost with minimal disruption to service. SLAC obtained and installed a beta unit and have been running it in production since September 2007.

The experiences described are with the Early Access version of the PBB. The production version of the box has engineering changes based in part on our experiences.

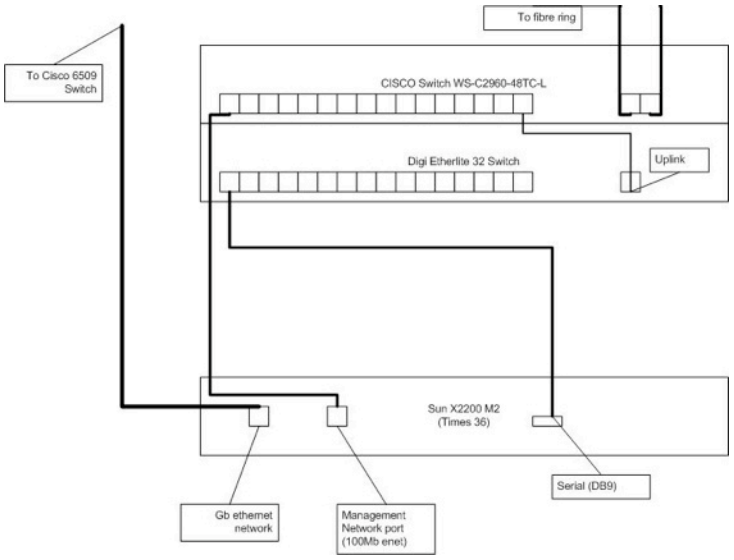
Planning

The rated capacity of a PBB is 200 KW of power and cooling. Our proposed load was only going to be about 70 KW, however we needed to account for possible future upgrades, and to add reserve capacity, to avoid operating at the rated limits of the supplies if we filled the box to capacity. We therefore planned for the acquisition of a 225 KW Outdoor Integrated Power Center (OIPC) to supply the power, and a 60 Ton chiller to supply the cooling needed.

For site selection, a number of locations at the Lab were considered. Alternative locations next to the main power substation, where power would be easily available, and next to the main campus chiller, where cold water would be easily available, were eliminated for operational reasons. Finally, a site behind the main computer center was chosen, next to the 4 MW substation that supplies power to the center. This would provide for easy access to power, and ready servicing of the box by center personnel.

To provide proper support for the PBB, the chiller, and the OIPC, three concrete pads were poured. The pads provide seismic stability for the boxes, anchor points for strapping the units down, and grounding rods.

We purchased the PBB along with 252 Sun X2200 M2 servers. These servers each had two dual-core AMD Opteron processors at 2.66 GHz, 8GB of memory, and two 500GB disks. Sun integrated the servers into the PBB at their plant in Oregon, wired them to the Cisco 6509 switch that SLAC provided, and performed burn-in testing. The PBB was then delivered to SLAC complete with payload.



Delivery, Connection, and Startup

On July 14, 2007, the PBB was delivered to SLAC. A crane lifted the 23,400 pound box from the flatbed truck onto the concrete pad. The only consequence of the shipping was that about half of the power cord pulled out of the servers due to the shaking on the truck. This was discovered and remedied on the startup date, as the racks were not pulled out for inspection until then.

On July 30, the chiller and OIPC were delivered to site and placed on their pads. Connection of the unit commenced about two weeks later, and finished in early September. While the PBB is designed for quick connections to generators and chillers for mobile deployment, we were doing a robust installation that was expected to be in production for a number of years.

Blackbox Delivery

Startup was on September 17. The power panel was energized, then the chiller was started. It was not initially planned to turn on the payload machines that day, however the engineers tuning the chiller needed a heat load for the chiller to work properly. The network switch was configured that day, and integrated with the SLAC network.

The following day the payload systems were configured for the SLAC network. A few systems needed to have power cords, cables, or memory reseated to bring them up. Because the systems had already had the operating systems installed offsite, configuration was straightforward. A configuration error in the network setup was readily corrected: the integration center had failed to set the base network address resulting in no default route for the systems. A working system on the same subnet was able to connect to each machine in turn and correct the setting. After that standard SLAC network configuration scripts took over.

During the time the PBB was on order and being installed, SLAC changed some details of its networking scheme. Once the systems were brought up and configured, we then needed to re-address and reboot the cluster, and change the switch to the new addressing scheme. This stage took another few days to complete.

After running a test load on the machines for the remaining days of that week, the machines were opened to batch work on September 25. In the time since then, their productivity and reliability has been equivalent to an identical set of machine housed in the computer center.

Service

Service procedures for the PBB are somewhat different from racks in the computer center. Procedures require two people to be present, in case of injury or entrapment. Both external doors at either end must be opened, both to provide two paths of exit from the box, and because one must often walk around the box to get at a rack being serviced from the other side. When closing the box, one end must be closed and secured first, and then a visual inspection

performed from the other door, to insure that no one is inside before locking up. We have also equipped the box with a telephone inside as a further precaution in case someone is stuck inside.

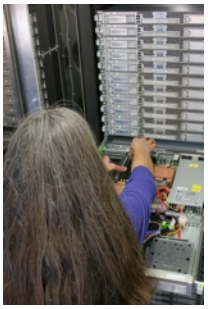
When servicing a system, first both outer doors are unlocked and propped open. These are heavy, and require a certain amount of precaution against back strain when opening. Then the inner doors are opened, two at one end and one at the other.

A special "slider" is used to roll the rack to be serviced into the aisle. First it is rolled down the aisle to the required rack. It has two sets of rollers at 90° to allow it to be rolled in either direction by lowering one set or the other.

Racks are anchored at the top by a pin and at the bottom by four bolts. Those must be removed before the rack can be slid out for service.

After that, the rack can be jacked up and rolled out into the center aisle for service. Systems can be slid out for service on their rails without powering down or affecting the other systems in the rack, similar to the procedures in racks inside the computer center.



Conclusions and Future Plans

The Sun Project Blackbox was a cost-effective and expeditious way to meet SLAC's immediate needs for additional computer equipment without costly and lengthy improvements to the computer center. The partnership with Sun on this project was beneficial, as Sun engineers worked closely with us to understand our requirements, followed our deployment of the box, and incorporated lessons learned back into the production product. The time to deploy was roughly six months from project kickoff to operation. With the knowledge gained, it could probably be reduced to three months for another such unit. In operation, it has been reliable so far, and the payload machines have behaved as expected. Service is somewhat cumbersome, but is fine for highly redundant systems. We tend to batch together several failed systems and schedule them all for service at the same time. Service would be more onerous for more critical systems, as the effort to open the box and the rack is significant.

It is currently anticipated that a second PBB will be the best way to answer BaBar's capacity needs for their next data-taking run. This study is currently underway.